Studying Large Language Model Generalization with Influence Functions

Year: 2023 | *Citations: 249* | *Authors: R. Grosse, Juhan Bae, Cem Anil*

Abstract

When trying to gain better visibility into a machine learning model in order to understand and mitigate the associated risks, a potentially valuable source of evidence is: which training examples most contribute to a given behavior? Influence functions aim to answer a counterfactual: how would the model's parameters (and hence its outputs) change if a given sequence were added to the training set? While influence functions have produced insights for small models, they are difficult to scale to large language models (LLMs) due to the difficulty of computing an inverse-Hessian-vector product (IHVP). We use the Eigenvalue-corrected Kronecker-Factored Approximate Curvature (EK-FAC) approximation to scale influence functions up to LLMs with up to 52 billion parameters. In our experiments, EK-FAC achieves similar accuracy to traditional influence function estimators despite the IHVP computation being orders of magnitude faster. We investigate two algorithmic techniques to reduce the cost of computing gradients of candidate training sequences: TF-IDF filtering and query batching. We use influence functions to investigate the generalization patterns of LLMs, including the sparsity of the influence patterns, increasing abstraction with scale, math and programming abilities, cross-lingual generalization, and role-playing behavior. Despite many apparently sophisticated forms of generalization, we identify a surprising limitation: influences decay to near-zero when the order of key phrases is flipped. Overall, influence functions give us a powerful new tool for studying the generalization properties of LLMs.